Haolong Chen

☎ 18316035908 · ✉ haolongchen1@link.cuhk.edu.cn · github.com/IfReasonable

Education

The Chinese University of Hong Kong, Shenzhen, Computer and Information Engineering, *PhD* 2028.6

- **Supervisors**: Guangxu Zhu, Tsung-Hui Chang

Jinan University, Software Engineering, *Bachelor* 2023.6

- **Supervisors**: Guanghua Yang, Xinyuan Zhang

Experience

Shenzhen Research Institute of Big Data, Research Assistant 2023.5 – Now

- Research in the fields of efficient inference for LLM, efficient training for LLM, spatio-temporal data analysis, and artificial intelligence in wireless communication.

Jinan University High-Performance Computer Team, Team Member 2021.6 – 2023.5

- Participate in international high-performance computer competition ASC21, SC21.

Publications

An overview of domain-specific foundation model: key technologies, applications and challenges

- **Haolong Chen**, Hanzhi Chen, Zijian Zhao, Kaifeng Han, Guangxu Zhu, Yichen Zhao, Ying Du, Wei Xu, Qingjiang Shi
- Accepted by SCIS (CCF A).

STM3: Mixture of Multiscale Mamba for Long-Term Spatio-Temporal Time-Series Prediction

- **Haolong Chen**, Liang Zhang, Zhengyuan Xin, Guangxu Zhu
- Submitted to ACM KDD.

AdaMeZO: Adam-Styled Zeroth-Order Optimizer for LLM Fine-tuning Without Memorizing the Moments

- Zhijie Cai*, **Haolong Chen***, Guangxu Zhu
- Submitted to NeurIPS.

FeedSign: Robust Full-parameter Federated Fine-tuning of Large Models with Extremely Low Communication Overhead of One Bit

- Zhijie Cai*, **Haolong Chen***, Guangxu Zhu
- Submitted to IEEE TMC (CCF A).

A Semi-Supervised Approach for Telecom Poor User Experience Root-Cause Classification

- Qizhe Li, **Haolong Chen**, Jiansheng Li, Shuqi Chai, Guangxu Zhu
- Preparing for journal submission.

First Token Probability Guided RAG for Telecom Question Answering

- Tingwei Chen, Jiayi Chen, Zijian Zhao, **Haolong Chen**, Liang Zhang, Guangxu Zhu

Projects

Efficient Collaboration of Multi-Agent | *LLM Inference, Multi-Agent Collaboration*

- Constructing an efficient multi-agent debate system from the perspective of information diversity.

Efficient Inference of LLMs in Edge Intelligence System | *LLM Inference, LLM Routing*

- Model the LLM routing workload scenario using economic auction models.

Efficient Inference of LLMs in Edge Intelligence System | *LLM Inference, Edge Intelligence*

- Developed a reasoning-enhanced edge inference framework that leverages cloud large-model reasoning outputs to build an RAG corpus, thereby enhancing edge small-model capabilities.

Efficient Training of LLMs in Edge Intelligence System | *LLM Training, Edge Intelligence*

- To address challenges of limited memory, slow computation on edge devices, and high communication overhead in federated learning, a gradient compression algorithm based on the random seed is proposed.
- A lightweight method is introduced to enable second-order momentum estimation by recording a small number of historical seeds, which significantly accelerates training while incurring minimal memory overhead.

Guangdong Major Project of Basic and Applied Basic Research: Research on Key Technologies of 6G Networks Enhanced by Environment | *Communication KPI Modeling*

- Charging of the sub-project on modeling and simulation of user spatiotemporal distribution and traffic flow, which is part of the larger project on spatiotemporal state modeling and simulation of network elements.
- Developed a joint spatiotemporal traffic modeling approach using multiple base stations and proposed a novel spatiotemporal traffic prediction model that integrates the Mamba long-term sequence neural network, dynamic graph convolutional network, and sparse mixture of experts.

National Key Research and Development Foundation: Learning Optimization Theory and Methods and Their Applications in 5G Networks | *Communication KPI Modeling*

- Responsible for user-side performance modeling based on spatiotemporal integration in Subproject: Performance Modeling of 5G Network Systems.
- Proposed a method for spatiotemporal user performance modeling based on a multimodal large language model, which can integrate time-series user performance data with text descriptions of the dataset and network environment information surrounding the service area to achieve high-precision prediction.

Multidimensional User Experience Modeling (Huawei - SRIBD) | *Communication KPI Modeling*

- Constructed a time-series classification model for user experience anomalies.
- Proposed a data augmentation method based on diffusion models to address the issue of overfitting due to the limited amount of labeled data.

Spectrum Efficiency Modeling with Measured MIMO Channel (Huawei - SRIBD) | *Communication KPI Modeling*

- Utilized real-world 5G MIMO measurement data from multiple grids and cells to predict record-level spectrum efficiency under multi-grid and multi-cell scenarios.

Reviewer for International Research Conferences

- NeurIPS 25, ICASSP 24, ICC 24,25, GLOBECOM 25, ICCC 25, WCNC 24,25, PIMRC 25, etc.

Patents

Method, Apparatus, Electronic Device, and Storage Medium for Traffic Prediction in Wireless Communication

- Inventors: **Haolong Chen**, Zhengyuan Xin, Guangxu Zhu, Assignee: Shenzhen Big Data Research Institute, Patent Number: ZL202511087894.7, Date of Authorization: 2025.10.28.

Predictive Method and Related Apparatus based on Multimodal Large Models for Communication Key Performance Index Prediction

- Inventors: **Haolong Chen**, Guangxu Zhu, Qingjiang Shi, Assignee: Shenzhen Big Data Research Institute, Patent Number: ZL202510542918.7, Date of Authorization: 2025.10.17.

Model Training Methods, Text Classification Methods, Devices, Electronic Devices, and Media

- Inventors: Zhijie Cai, **Haolong Chen**, Guangxu Zhu, Qingjiang Shi, Assignee: Shenzhen Big Data Research Institute, Patent Number: ZL2025103509755, Date of Authorization: 2025.9.16.

Communication and Memory Efficient Distributed Training Methods for Large Models and Text Classification Methods

- Inventors: Zhijie Cai, **Haolong Chen**, Guangxu Zhu, Assignee: Shenzhen Big Data Research Institute, Patent Number: ZL2025100670425, Date of Authorization: 2025.8.12.

Predictive Method, Apparatus, Electronic Device, and Storage Medium for Spectrum Efficiency

- Inventors: **Haolong Chen**, Guangxu Zhu, Qingjiang Shi, Assignee: Shenzhen Big Data Research Institute, Patent Number: ZL2023115716969, Date of Authorization: 2024.2.23.

Software Monographs

Semi-Supervised Training and Solution System for Spectral Efficiency Prediction Algorithms Based on Large-Scale User Measurement Report Data v1.0

- Assignee: Shenzhen Big Data Research Institute, Assignment Number: 2024SR1450315, Date of Authorization: 2024.9.29.

Skills

Programming

- Proficient in: Python, PyTorch, Linux
- Familiar with: Matlab, C/C++, MySQL, Git, Java, Web Frontend Development, Web Backend Development, TensorFlow

Languages

- English (IELTS: 6.5, CET-4: 548, CET-6: 542)
- Chinese (mother tongue)